Linguistic Variation in Greek Papyri: Towards a New Tool for Quantitative Study

Mark Depauw and Joanne Stolk

The digital revolution and papyrus linguistics

The growing digitization of classics has put splendid tools at our disposal, many of which have transformed our daily scholarly practice. For Greek papyrology, research without the *Papyrological Navigator* [PN], which combines i.a. the full text of the *Duke Databank of Documentary Papyri* [DDbDP] and the metadata of the *Heidelberger Gesamtverzeichnis* [HGV],[1] is almost unimaginable today. Finding parallels for words or expressions in the pre-digital era used to be a matter of thorough—and rare—human expertise or the result of a painstaking and time-consuming search through all extant sources or their indices, sometimes with the help of dictionaries or concordances, assuming of course that a complete library with text editions was at hand. Digital tools have now put most types of heuristics at the disposal of everyone with an Internet connection, reducing the investment of time to a fraction of what it was before.

The digital revolution has thus greatly facilitated what we have always done. Yet the change may be more fundamental.

[1] The PN (www.papyri.info) combines the full text of the DDbDP (which no longer has its own separate user interface) and the metadata of the HGV (also accessible at www.rzuser.uni-heidelberg.de/~gv0/). It also includes information from the *Advanced Papyrological Information System* [APIS] (the main gate of access of which is now also the PN), the *Bibliographie Papyrologique*, and other projects that are less relevant in this context. It cooperates closely with *Trismegistos* [TM] (www.trismegistos.org), a platform for the study of texts from Egypt between 800 B.C. and A.D. 800.

Greek, Roman, and Byzantine Studies 55 (2015) 196–220

Traditional scholarship often relies on connoisseurship. Experts state their opinion on the basis of an intimate knowledge of relevant sources. Digitization now increasingly makes it possible and plausible to quantify many of these traditional expert opinions. Research questions, which exclusively relied on connoisseurship before, can be tackled in a quantitative way. Of course, thorough historical and philological background knowledge and an intimate acquaintance with the source material remain essential, but statements can now more easily be evidence-based, rather than impressionistic.

The Trismegistos platform has over the last couple of years been exploring long-term onomastic evolutions through quantification.[2] Yet there are many other areas where similar methods can be productive. One of these is the study of linguistic developments, where for modern languages corpus linguistics are now the basic starting point. Yet, although the full text is at hand, the potential of papyri remains largely untapped. Recent approaches to their language focus on diversity, language contact, and language variation and change.[3] Although Gignac's grammar offers a good starting point for the study of phonological and morphological variation,[4] a book cannot provide an exhaustive and dynamic overview of the attested variants. In order to fill the void in the study of

[2] See e.g. M. Depauw and W. Clarysse, "How Christian was Fourth Century Egypt? Onomastic Perspectives on Conversion," *VigChr* 67 (2013) 407–435; M. Depauw and B. Van Beek, "People in Greek Documentary Papyri: First Results of a Research Project," *JJP* 39 (2009) 31–47.

[3] See for example the papers in T. V. Evans and D. D. Obbink (eds.), *The Language of the Papyri* (Oxford 2010); A. Mullen and P. James (eds.), *Multilingualism in the Graeco-Roman Worlds* (Cambridge 2012); M. Leiwo et al. (eds.), *Variation and Change in Greek and Latin* (Helsinki 2012). The study of social and regional variation and the role of language contact is well established for the Latin language by the work of J. N. Adams: *Bilingualism and the Latin Language* (Cambridge 2003), *The Regional Diversification of Latin 200 B.C.–A.D. 600* (Cambridge 2007), *Social Variation and the Latin Language* (Cambridge 2013).

[4] F. T. Gignac, *A Grammar of the Greek Papyri of the Roman and Byzantine Periods* I–II (Milan 1976).

(morpho)syntactic change, a morphologically and syntactically annotated database of all digitally available papyri is highly desirable, but this is a long-term project. The study of language change starts with language variation, and in fact a large portion of orthographic, morphological, and sometimes even morphosyntactic variants is already encoded in the online text available through the PN. This paper is an introduction to the development of a tool that uses these encoded variants in digital texts. The annotations may be a stepping stone towards more sophisticated analyses of variation and change in the language of the papyri.

1. *Encoding variation in (digital) Greek papyrus texts*[5]

The PN now includes the full text of each published Greek papyrus, ostracon, or other papyrological document.[6] In a long and complicated process, the original beta-code digitized text of the pioneering DDbDP was transformed to Unicode with annotations in *Extensible Markup Language* [XML].[7] This has made the PN a very powerful tool for lexical searches, yet the user interface currently does not allow searches for specific XML-markup. Through the generous Open Access license of the PN it is possible, however, to download the most recent version of the Unicode/XML-annotated text or to scrape the *HyperText Markup Language* [HTML] visible in the browser from the individual pages for each record, as we have done.

The XML annotations in the PN's full text are in *Text Encoding Initiative* [TEI]-compliant EPIDOC, a particular 'flavor' of XML which has quickly become the standard for marking

[5] We here omit the Latin and Coptic corrections, the former because it has always been a minority language in documentary papyrology, the latter because only a few texts are currently included, although coverage is on the rise.

[6] For literary papyri a new tool called *Digital Corpus of Literary Papyri* is currently being developed as a counterpart and complement to the PN: see www.neh.gov/divisions/odh/grant-news/announcing-4-nehdfg-bilateral-digital-humanities-program-awards.

[7] See papyri.info/docs/ddbdp.

up ancient texts.[8] The system thus allows annotating specific passages as lost, abbreviated, damaged but still legible, etc.

In EPIDOC-XML it is also possible to integrate statements that analyze the actual text found in the papyrus as a variant of a form which is more 'mainstream' in the Greek language. Traditionally, these editorial interventions are referred to as editorial 'corrections', which gives the impression that the editor is correcting the language of the writer of the papyrus. Modern linguists are of course not necessarily interested in 'correcting' the writers, but in the linguistic analysis of synchronic and diachronic variation. However, if deviating forms are not annotated in any way, it becomes almost impossible to find them in a digital environment. For example, if one is interested in the use of the accusative pronoun με, including the variant spellings of the form, the almost 200 attestations of μαι can be found easily through a separate search. It will be a tedious job, however, to filter out the cases where με is spelled μη from the more than 7000 attestations of the conjunction μή. Variant spellings of longer, more complicated words or examples of more abstract linguistic phenomena may even turn out to be impossible to find through traditional searching methods, using exact spellings of concrete lexical items. The editorial practice to comment on forms that are unexpected from a lexical or grammatical perspective is therefore useful not only for the analysis of individual words attested in variant spellings, but also for the study of specific linguistic phenomena in large digital text corpora.

In the absence of more detailed linguistic annotations, the long tradition of editorial comments on linguistic forms can and should therefore be exploited for linguistic purposes.[9]

[8] See en.wikipedia.org/wiki/EpiDoc or e.g. Gabriel Bodard, "EpiDoc: Epigraphic Documents in XML for Publication and Interchange," in F. Feraudi-Gruénais (ed.), *Latin on Stone: Epigraphic Research and Electronic Archives* (Lanham 2010) 101–118.

[9] For practical reasons some of the traditional terminology involving e.g. 'corrections', 'errors', 'irregularities', 'regularization', etc. is still employed in

There are several ways in which these types of comments are integrated in the PN. First, it is possible to annotate the word using a regularization tag or a correction tag. The PN then shows the original form as it is written on the papyrus in the main text, but with an asterisk referring to the critical apparatus at the bottom where the 'regularized' or 'corrected' form is provided. Other possible text-critical annotations relevant here are editorial additions of missing letters or words, and editorial deletions of superfluous text, marked by in-text pointed brackets < > or accolades { } respectively, embracing the omitted or superfluous text, as in the traditional Leiden system:[10]

Editor's action	HTML main text	HTML apparatus	underlying XML
Regularization	*original version**	l. *regularized version*	<reg>*regularized version*</reg> <orig>*original version*</orig>
Correction	*original version**	l. *corrected version*	<corr>*corrected version*</corr> <sic>*original version*</sic>
Addition	<*omitted text*>	–	<supplied reason="omitted">*omitted text*</supplied>
Deletion	{*superfluous text*}	–	<surplus>*superfluous text* </surplus>

Of course this set of tags allows a fair amount of variation in annotation. Although the regularizations, corrections, and the use of brackets in the Leiden system all denote an editorial comment, they are often considered to reflect different situations: the correction tag or brackets indicate simple scribal

this paper as well as in the database.

[10] Ancient corrections, such as secondary-stage additions (indicated with \ / in the text following the Leiden transliteration system; XML <add place="above"> </add>) or in-text corrections ("corr. ex" in the apparatus; in XML a combination of <add place="inline"> </add> and <del rend="corrected"> </del>) are not taken into account here, although they may also be relevant and very interesting to study for linguistic reasons. We will expand the Trismegistos Text Irregularities database with these 'ancient scribal interventions' in the not too distant future.

mistakes, whereas the regularization tag points to the 'normalization' of morphological variation or irregular spelling. In practice, however, these distinctions are not always applied consistently. For a case of itacism such as χιρός instead of χειρός, the editor (both in print or in an online environment) can opt to transcribe χ<ε>ιρός using the addition tags, or he can alternatively employ the correction tags resulting in χιρός(*) in the main text and χειρός in the critical apparatus.[11] Although it is obvious to a human that these are basically the same thing, the distinction may have repercussions in a digital environment, as we shall see.

2. *Collecting variants in a database*

Variations may be annoying for those interested in the contents of the papyrus documents because they hamper legibility and complicate digital searches, but they are interesting for linguists. They may show imperfect knowledge of an acquired second language, or they may illustrate changes in the language.[12] Of course, it is a slippery slope to decide where common variants become irregular enough for regularization and when scribal errors require correction, and different editors make different choices, in the past as well as today. But a database with editors' corrections and regularizations could be an important first step towards a more flexible study of linguistic variation in Greek papyri. A digital tool could allow permanently up-to-date versions of exempla for the many

[11] Of course the editor can also decide not to correct the irregular form, because it is not considered irregular enough to warrant attention: see §4 below on the role of editorial practice.

[12] See P. Fewster, "Bilingualism in Roman Egypt," in J. N. Adams et al. (eds.), *Bilingualism in Ancient Society: Language Contact and the Written Word* (Oxford 2002) 220–246, esp. 232–236. For examples of imperfect learning of a second language see M. Vierros, *Bilingual Notaries in Hellenistic Egypt: A Study of Greek as a Second Language* (Brussels 2012), and for the study of language change in the papyri e.g. P. James, "Variation in Complementation to Impersonal *verba declarandi* in Greek Papyri from the Roman and Byzantine Periods," in *The Language of the Papyri* 140–155.

phonological phenomena described by e.g. Mayser's grammars, Teodorsson's study of Ptolemaic phonology, or Gignac's grammar of the Roman and Byzantine periods, as well as providing a start for the analysis of morphosyntactic variation.[13] A dynamic overview of editorial corrections from the past and present could also be used as a tool to explore the varying modern responses to linguistic variation and may be helpful to develop new guidelines for editorial practices better suited for modern linguistic studies.

Prompted by an email from the second author of this paper about her ongoing research on case interchange in Greek papyri, the first author decided to mine the texts in the PN for relevant annotations (state of 4 Jan. 2014). For practical reasons the HTML was used rather than the underlying XML, which should be the basis for improved later versions. The process consisted of a scrape of the DDbDP full text (.html) on the basis of Trismegistos numbers (texid), followed by a conversion to plain text form (.txt), and an import in Filemaker 13 for data manipulation. The letters and words marked with asterisks in the main text were then matched to the corresponding entries in the critical apparatus and checked; the apparatus entries and in-text annotations were moved to a separate related database; editorial regularizations, corrections, additions, and deletions were separated from other markup, which was removed where irrelevant or undesired; and finally textual metadata such as provenance and date were drawn in from HGV and Trismegistos. This database was then exported to MySQL and a user interface was set up at www.trismegistos.org/textirregularities, allowing free access to all interested users.

[13] E. Mayser, *Grammatik der griechischen Papyri aus der Ptolemäerzeit* (Berlin 1926–1938); S.-T. Teodorsson, *The Phonology of Ptolemaic Koine* (Göteborg 1977; see the review by W. Clarysse in *BibO* 40 [1983] 81–86 for possible pitfalls in the creation of lists of attestations); and Gignac, *Grammar*. For an introduction to Greek linguistic evolutions in the papyri see E. Dickey, "The Greek and Latin Languages in the Papyri," in R. S. Bagnall (ed.), *The Oxford Handbook of Papyrology* (Oxford 2009) 149–169, with more literature.

3. *First results*

This database is in many ways imperfect, but nevertheless it allows a first innovative quantitative approach to the subject of variation in the Greek language as written in Egypt.[14] We will first look at the chronological developments, after which we will have a closer look at the type of editors' corrections found in the database and at the possible variables governing chronological variation.

3.1. The chronological evolution of editorial corrections

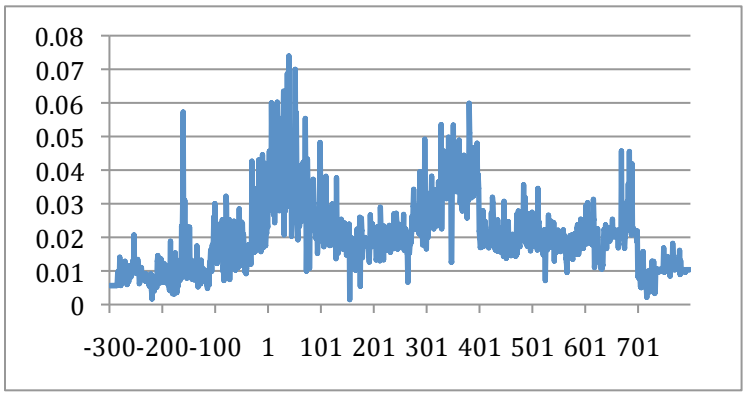
The first quantitative question that a database of DDbDP editors' corrections can answer is how many words the linguistic corpus of Greek papyri contains and how many words on average were corrected in each text (both in absolute and relative terms). The number of words for each text can be calculated on the basis of the number of spaces: from a single word for texts such as *SB* XVIII 13938 (TM 25390) to 31,961 for *P.Sorb.* II 69 (TM 20110). In all, the 52,756 papyrological texts collected from the PN contained 6,558,982 words (average word length = 124.33, median = 49). In a next step the number of editorial regularizations and corrections can be calculated, both those in the apparatus (121,088), and the in-text additions (6921) and deletions (3052), in all 131,061. This means that almost exactly 2% of all words in Greek papyri are the subject of editorial intervention of this kind.

Yet, as could be expected, this general figure conceals a great diversity. Almost half of all papyri (25,730 or 48.8%) do not include even a single correction, and some of these 'faultless' texts are even up to 3109 words long, e.g. *P.Lond.* I pp. 140–149 no. 119 (TM 20001). Others are rife with 'errors' and 'irregularities' of all kinds, with a case like *SB* XIV 12030 (TM

[14] Obviously, the database can only be based on the written language of the papyri. Although this does not necessarily reflect the changes in the spoken vernacular, careful analysis of the texts can provide a deeper insight into the underlying factors: cf. M. Leiwo, "Introduction: Variation with Multiple Faces," in *Variation and Change* 1–11, esp. 3–5.

34811) at the top, with 34 out of 69 words or 49.2% corrected. The average percentage of corrected words per text is 2.3%, slightly higher than the average calculated on the total number of words, but on the whole not divergent enough to warrant the presumption that a few very long texts with very few corrections distort the figures.

The second question that arises is whether there was any chronological development in the number of corrections, based on the dates of the papyri. To investigate this, we created a 'weighted dates' graph in which the number of corrections and the number of words in a papyrus are spread out over the time range covered by the papyrus.[15] In such a graph a papyrus dated between 299 and 200 B.C., counting 40 corrections for 400 words, is counted as 0.4 corrected words (40/100) for 4 words (400/100) in each of the hundred years from 299 to 200 B.C. This results in Graph 1.



Graph 1. Percentage of corrected words
in Greek papyrological texts, 300 B.C. to A.D. 800

The graph shows some clear longer-term evolutions as well as

[15] For a detailed description of the procedure see B. Van Beek and M. Depauw, "Quantifying Imprecisely Dated Sources: A New Inclusive Method for Charting Diachronic Change in Graeco-Roman Egypt," in *Ancient Society* 43 (2013) 101–114.

some interesting highs and lows. First of all the Ptolemaic period in general has much lower percentages, around 1% or less, until about 110 B.C. The exception is a brief period, between 164 and 156 B.C., when there are higher figures, up to 5.7% in 161 B.C. From the very end of the second century B.C. onwards, the average number of irregularities increases, which particularly at the very end of the first century B.C. and in the first half of the first century A.D. leads to peaks much above the average. In the second half of the first century A.D. and the second century, figures drop and stabilize around 2% until the early fourth century, when an upward trend again emerges that will last until the end of the century. In the fifth century there is a return to 'normality', lasting until the eighth century, when the number of corrections returns to the level of Ptolemaic times (but based on far fewer sources and thus perhaps not statistically significant).

3.2. Type of irregularity

Editors correct all kinds of things in the text. Some are factual errors, e.g. a mistake in the name of the father of a well-known individual or a calculation error in an account; some are graphic errors, e.g. haplography, dittography, or inversion;[16] others are morphological or morphosyntactic mistakes, e.g. the use of a nominative case instead of the expected genitive or an ending from a different paradigm; and yet others are orthographic irregularities, often caused by phonological changes resulting in phonetic similarity.[17] To distinguish between these types of the more than 100,000 editorial corrections is obviously a major task, which is only in its beginning stages. As stated above, in the Duke XML a distinction is fore-

[16] See e.g. the categorization of mechanical errors in Gignac, *Grammar* I 59.

[17] Semantic and syntactic mistakes are sometimes corrected by the scribe in antiquity, but usually not by modern editors in the text or apparatus. Additional remarks of a lexical, semantic, syntactic, or pragmatic nature can sometimes be found in the editorial commentary on the papyrus.

seen between <reg> and <orig> tags for regularizations on the one hand, and <corr> and <sic> tags for errors on the other. But the distinctions between the two need clarification through further discussion, and for historical reasons most editorial interventions have been marked as regularizations.

We have therefore developed a computer-assisted way to describe the difference between what was written in the ancient text and the editor's regularized or corrected version in the 121,088 apparatus entries.[18] The procedure is based on a comparison of the original form and the corrected version, making abstraction of the 'noise' caused by brackets, Greek accents, and further diacritic signs. Starting from the beginning of the two 'cleaned' words, the letters are compared and the position of the first diverging letter is calculated. After further manipulation through replacement, the computer then suggests an interpretation of the difference between the two versions. A human interpreter then accepts or rejects the suggestion, and by subjecting the set to repetitive slightly adapted algorithms, an increasing number of interpretations can be reached. Obviously, the more similar the two versions are, the easier the task is, and this should be taken into account when interpreting Table 1 with a list of the most common irregularities.[19]

The fifteen most common irregularities account for 56,791 apparatus entries or 53% of the 106,589 items currently processed. A power law distribution with a limited number of types accounting for a large majority of all observations and a long tail is indeed expected in a frequency ranking. The irregularities accounted for also largely confirm our expectations based on previous research. This means that the editors' decision to correct a form might have been influenced by the analysis found in the grammars, but also that high frequency has not prevented them from marking 'obvious' variations,

[18] We have omitted the in-text additions and omissions in a first instance.

[19] The principles for the description of these types of irregularities are discussed in more detail in §5.1 below.

such as ι/ει interchanges.[20] Interestingly, many of the entries in the list are mirror observations of related phonetic phenomena: monophthongisation, resulting in ι/ει variation or ε/αι alternation;[21] loss of vowel length distinction, e.g. ο/ω confusion; and problems with consonantic voice, e.g. γ instead of κ.

Irregularity	*Frequency*
ι instead of ει	17604
ει instead of ι	11540
ο instead of ω	5644
ω instead of ο	4578
ε instead of αι	3549
ωι instead of ω[22]	1735
αι instead of ε	1635
omission of ο	1565
omission of ν	1564
omission of ς	1427
ε instead of α	1420
υ instead of οι	1305
γ instead of κ	1269
ου instead of ω	1167
τ instead of δ	969

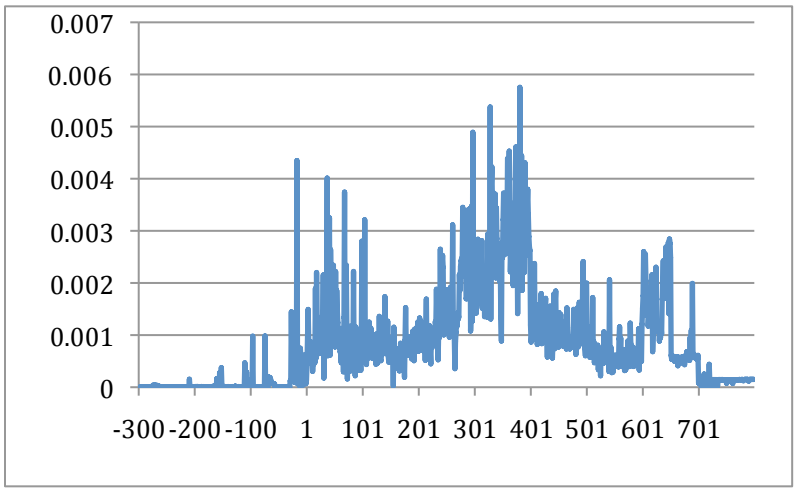
Table 1. The fifteen most common irregularities (state 27 Feb 2014)

[20] Of course many examples of iotacism will have been left unmarked, making several phenomena even more frequent than reflected in Table 1.

[21] Interestingly, other forms of itacism involving η are somewhat less common, e.g. ι instead of η with (846 exx.), η instead of ει (767), η instead of ι (460) and ει instead of η (436). This may confirm the later date of this η/ι merger (cf. Dickey, in *Oxford Handbook* 152).

[22] The reverse phenomenon (ω instead of ωι) is much rarer (only 9 exx.!) because this is normally silently corrected by the editors in the form of *iota subscriptum*. In fact it should be discussed at some later stage whether this should not be annotated as an editorial correction also.

In order to observe the chronological developments in more detail, we can take a close look at one of the examples of monophthongisation (Graph 2) on the basis of ε/αι interchange rather than ι/ει itacism.[23] Again we have set out the number of attestations of this type of editorial intervention against the number of words, in a weighed date graph.[24]



Graph 2. Percentage of ε/αι interchange in Greek papyrological texts, 300 B.C. to A.D. 800

There are differences between Graph 1 with the general evolution and Graph 2, but on the whole there are more similarities. The Ptolemaic period is again underrepresented, this time to the extent of being almost invisible. Only at the very end of the first century B.C. does the number of attestations rise gradually,

[23] Not only because ι/ει itacism is more likely to be ignored by editors, but also because it can alternatively be expressed outside the apparatus, in the form of the omission or deletion of the ε through in-text sharp brackets or accolades.

[24] Although it would probably be more correct to compare with the number of words in which αι or ε are present, the pattern would in all likelihood be very similar in view of the high frequency of both.

to reach a peak at the end of the fourth century A.D. Then percentages go down again, with a similar low for the fifth and sixth centuries. In the seventh century there seems to be a rise at first, but then a return to 'normal' levels. The eighth century again shows levels comparable to those of the Ptolemaic period. The chronological distribution is of course to some extent related to the process of monophthongisation. But other factors than changes in pronunciation might explain the evolving frequency of linguistic phenomena in writing or their correction by the editors, and we turn our attention to them now.

3.3. Variables governing chronological variation

The frequency of editorial corrections thus clearly fluctuates over time, but the next question is how to interpret these fluctuations. For some peak moments, very precise reasons can be established. Thus the brief rise in the number of corrections (Graph 1) around the middle of the second century B.C. is clearly caused by the Katochoi archive,[25] where two Egyptian-style eremites, Ptolemaios and his younger brother Apollonios, have produced many Greek documents that were not always written in standard language, to say the least. For longer-term evolutions, however, there might be many factors playing a role. For the study of a particular (socio)linguistic phenomenon it is common to distinguish a dependent variable (the type of irregularity under study) and several independent variables, such as time period, document type and provenance, or social background and native language of the writer.[26] Those variables are

[25] See B. Legras, *Les reclus grecs du Sarapieion de Memphis. Une enquête sur l'hellénisme égyptien* (Studia Hellenistica 49 [Leuven 2011]).

[26] For the amount of deviations in a document as related to the document type (register) and the scribe's linguistic background, including level of education and native language, cf. the distinction of user- and use-related variation in Leiwo, in *Variation and Change* 2. Especially the level of education is suspected to cause deviations, see T. V. Evans, "Linguistic and Stylistic Variation in the Zenon Archive," in *Variation and Change* 25–42 (esp. 40, where he concludes that the level of education rather than ethnicity would be the primary cause of nonstandard Greek); and K. Versteegh, "The

also relevant to describe the relation between the amount of corrections and historical developments. The rise in the early Roman period, for example, could be the result of restrictions on the use of local languages such as Demotic in an official context,[27] effectively banning Egyptian from public life and also from daily use in letters, petitions, or even oracle questions. This may have caused an influx of native Egyptian speakers writing in Greek, the effects of which could then have worn off in the second century A.D. and later. But this second-language hypothesis can hardly explain the new increase in irregularities in the fourth century. Alternative (or additional) explanations include general changes in language education of ancient scribes, the type of document that happens to be preserved and selected for publication, or other editorial practices.[28]

The type of text (private, public) and the genre (letter, account etc.) may influence the number of text irregularities. Accounts and lists are often written by professional scribes, and in contracts and other official documents one equally expects fewer mistakes because of the public and formal character, and the use of set phrases. Official documents might also contain more numbers and abbreviations that lower the relative frequency of linguistic irregularities. The type of text preserved may play an important role, for instance, in the drop in irregularities in very late texts, where the average is based on few documents, most of them written by well-trained scribes.

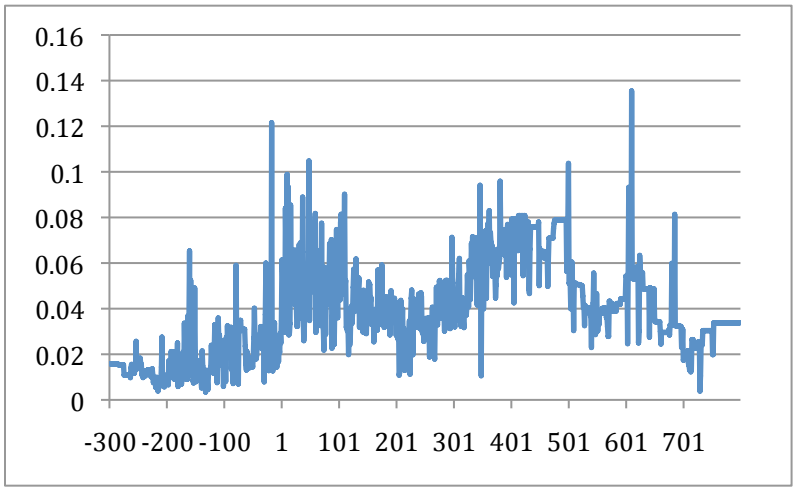
However, the text type/genre is difficult to monitor, since neither TM nor the PN (through HGV) have developed a fully standardized text typology. Fortunately, Delphine Nachtergaele, a Ph.D. student of the University of Ghent, is working on

Status of the Standard Language," in *Bilingualism in Ancient Society* 69.

[27] M. Depauw, "Language Use, Literacy, and Bilingualism," in C. Riggs (ed.), *Oxford Handbook of Roman Egypt* (Oxford 2012) 493–506.

[28] Perhaps the interest in early Christianity made editors select relatively more private letters from the fourth century for publication. For the impact of editorial practice on the spread of textual corrections see §4 below.

linguistic features of Greek letters.[29] Since she has done her database work inside TM, this has allowed us to use a set of 8311 letters where she has identified at least one epistolary formulaic element (Graph 3).



Graph 3. Percentage of corrected words in letters on papyri or ostraca, 300 B.C. to A.D. 800

The first and striking difference between Graph 3 and the previous ones is the much higher base level of text irregularities, with percentages that are often double those of the general Graph 1. This might confirm the general notion that substandard language can often be found in private letters.[30] The scribes of letters were not always professionals but often private individuals whose epistolary Greek was obviously far less classical and contained more irregularities than that used in

[29] See D. Nachtergaele, "The Asklepiades and the Athenodoros Archives: A Case Study of a Linguistic Approach to Papyrus Letters," *GRBS* 53 (2013) 269–293.

[30] Cf. E. W. Schneider, "Investigating Variation and Change in Written Documents," in J. K. Chambers et al. (eds.), *The Handbook of Language Variation and Change* (Oxford 2002) 67–96.

other genres.[31] Nevertheless the global chronological evolution remains similar, with lower percentages for the Ptolemaic period, a gradual increase in the first century A.D. and a stabilization at lower levels in the second century, followed by a second increase in the course of the fourth century. Only the drop in the fifth century is much less conspicuous for letters, although this figure is based on very few well-dated letters and thus perhaps subject to change. As the text type and genre both relate to the function and characteristics of the document and the level of professionalism and education of the scribe, these are important factors in explaining variation. A standardized typology of document categories is therefore highly desirable for the study of variation and change by means of the database.

It is clear that the chronological patterns that can be generated by means of the database are subject to the influences of many different factors and more study is necessary to allow for a quantitative approach to the assessment of language variation and change in Greek papyri. One of the variables that is most prominent in the database at this stage has to be the influence of the editors.

4. *The influence of editorial practice*

Finally, since the database is one of editors' interventions, some evolutions may at least partially be caused by variation of editorial practice at various levels. The attitude of editors towards the Greek they find in their papyri may for example vary individually, as rare cases of multiple editions of the same text in the PN show.[32] One also wonders whether the editors' annotations and regularizations have always been implemented

[31] It would yield an interesting study to analyze the occurrence of irregularities in different text genres in papyri. The use of formulae in official documents might turn out, just like the epistolary formulaic phrases, not to be a guarantee for standard language.

[32] E.g. TM 21653 (http://papyri.info/trismegistos/21653) for which in the PN two editions are currently provided: *SB* XXII 15614 and *P.Sel.Warga* 8. The former has two corrections that do not appear in the latter, the latter has one that does not appear in the former, one they have in common.

in the digital version, and if so, in what way. It is certainly an advantage that the digital format allows adding editorial corrections to the text at a later stage. This should make editorial practice more uniform, but in the absence of clear guidelines it could have the reverse effect and lead to a proliferation of comments without careful analysis of individual texts. This becomes even more important as texts increasingly start leading a public life in the PN without ever having been edited elsewhere.

There is therefore an urgent need for editorial rules specifying what type of irregularity should be corrected and what would be a suitable basis for comparison of substandard language. Scholars studying Byzantine Greek may be reluctant to correct non-classical orthographies, while these irregularities would for the Ptolemaic period no doubt have been the object of editorial intervention. Similarly, morphosyntactic variation such as the use of a genitive instead of the expected dative may be marked as 'errors' in the Ptolemaic or early Roman period, but some editors may well accept this as the standard form by Byzantine times, and may even mark the use of the 'archaic' dative as an irregularity, taking contemporary parallels as the basis for comparison. A case in point is the verb ὑπάρχω, 'belong to', which takes the dative in Classical Greek to express the possessor. In a papyrus of A.D. 374, the genitive μου was corrected by the editor to dative μοι,[33] because the dative is expected in Classical Greek and also seems to be the standard usage in the fourth century. However, the editors of a text of 507 decided to correct a dative to a genitive in the same construction,[34] probably because they believed this to be the standard expression in the sixth century.

In the end one should wonder whether Greek *koine* based on fifth century B.C. Attic should be the standard for comparison with the language found in documents more than a thousand years later. Perhaps it should, because the scribes themselves

[33] *BGU* XIII 2332.20 (TM 9723).

[34] *SB* XVIII 13947.15 (TM 18388).

modeled their language (especially the orthography) on that ideal, but there are also clear *koine* features that may require another treatment: a case in point is the spelling γίνεται for γίγνεται, which is common in *koine* and therefore usually not corrected by editors. Especially for (morpho)syntactic and pragmatic variation, careful discussion will be needed to determine what constitutes nonstandard usage and whether this needs to be marked up or not. For the study of substandard language in the Zenon archive, Evans suggested the usage of contemporary texts from the archive as a proper basis of comparison of the linguistic features.[35] However, the examples of ὑπάρχω with dative or genitive show that comparison to papyri from the same period might not be enough to avoid confusion, as the dative and genitive complements are both attested during the fourth to sixth centuries A.D. Therefore, contemporary parallels should not only date from the same period, but should also have the same provenance and contain the same linguistic construction in the same type of document in order to interpret phraseological variation correctly.[36] Even though detailed study of phraseological variation in papyri is important, this level of variation might be too complex to be dealt with in the apparatus of a text edition. The database could play a role in this discussion on the standard language and help to develop new principles for the editorial comments on linguistic variants. Its future extension towards scribal corrections could be particularly helpful for the question of standard and nonstandard language.[37]

[35] T. V. Evans, "Standard Koine Greek in Third Century BC Papyri," in *Pap.Congr* XXV (Ann Arbor 2010) 197–206.

[36] This issue will be addressed in a future article by the second author of this paper.

[37] This would give an idea of what the scribes themselves thought needed correction. However, various motivations for scribal correction can be identified, see for example R. Luiselli, "Authorial Revision of Linguistic Style in Greek Papyrus Letters and Petitions (AD i–iv)," in *The Language of the Papyri* 71–96.

Until clear rules emerge, one may wonder how representative editorial corrections are for the actual linguistic variation found in the papyri. Although the frequencies of the interchanges in Table 1, for example, do not seem very different from what we would expect, the database cannot be assumed to be exhaustive. For an in-depth study of the phenomena, linguistic analysis of all examples will need to be combined with careful testing of the representativeness of the overall results. For morphosyntactic features, we can estimate the coverage by comparing the editorial corrections with such a detailed study by the second author.[38] The editorial interventions involving the interchange of μου and μοι turned out to cover approximately 50% of all cases where μου might have been be used in a dative-like semantic role. This may seem unimpressive, but the interchange of case forms cannot be stated with certainty in all of these cases. The editors corrected almost all of the clear examples, but for obvious reasons they left out the more complex or ambiguous ones that can only be identified through linguistic analysis, comparison, and argumentation.[39]

Certainly in its current state, a database of editorial corrections clearly does not replace thorough analysis based on all attestations of a linguistic phenomenon. It does, however, offer access to those phenomena which are difficult to search for in the PN, e.g. (unanticipated) orthographic variants of a particular word, attestations of a particular phonological interchange, or the linguistic context of variation in the use of morphosyntactic categories. Without a linguistically annotated database for the papyri, these phenomena cannot be studied otherwise,

[38] J. V. Stolk, "Dative by Genitive Replacement in the Greek Language of the Papyri: A Diachronic Account of Case Semantics," *Journal of Greek Linguistics* (forthcoming).

[39] J. Humbert, *La disparition du datif en grec (du I^er^ au X^e^ siècle)* (Paris 1930) 171, mentions one of these ambiguous examples (*BGU* II 602.5–6). As the interpretation of a dative and a genitive are both possible in this text, the case form is not corrected by the editor even though this overlap of dative and genitive might be interesting for linguistic research.

except of course 'manually' in smaller sub-corpora. This does not imply, of course, that the phenomena mentioned in this paper have not been dealt with in previous studies, especially those related to phonology and morphology. Rather than replacing existing grammatical treatises with their scholarly interpretations, this new database seeks to generate more extensive and up-to-date lists of attestations, linked to the dynamic digital corpus in the PN. Most of the editorial corrections seem to be focussed on phonology and morphology rather than syntax, but some also deal with case forms and verbal conjugations. By taking some of the more intriguing editorial corrections as a starting point, interesting questions for future research and topics in need of linguistic analysis may be discovered that might have been missed out on in lexical searches or the study of individual papyri only.

5. *Setting up a cooperative environment for the database*

Defining and explaining linguistic variation may not always be straightforward. Nevertheless the collection of nonstandard forms in a digital environment remains an interesting first step towards tapping the linguistic potential of digitalized Greek papyrological texts. To maximize the potential of this tool, rules need to be established to describe the type of irregularities, to discover the types of scribal variation, and to try to overcome the bias of the traditional editorial corrections in the future. More importantly, however, flexible interaction between the full text curator (currently the PN) and the external text irregularities database needs to be assured.

5.1 Establishing rules for the description of irregularities

Ideally the difference between an irregular attestation and the corrected counterpart is described in an objective way, without interfering with the linguistic or diachronic interpretation of the phenomenon it illustrates. Therefore the description needs to be based on the smallest units represented: the phonemes and their graphic realizations. On the other hand, it should be possible to retrace available examples of the linguistic processes studied.

For graphic errors and orthographic variation due to

phonetic similarity, the description is reduced to three simple actions: the interchange, omission, and addition of signs (Table 2). For a meaningful interpretation, the position of the irregularity in the word is relevant as well. Separate fields will provide information such as initial, medial, or final, and 'before α' and 'after α'.

Linguistic process	*Description*	*Example*
Interchange	α instead of β	ει instead of ι (initial): ειναι
Omission	omission of α	omission of ς (final): μητρο
Addition	addition of α	addition of ν (final): πατεραν
Metathesis/inversion	αβ instead of βα	ορ instead of ρο (medial): Κορκ-
Haplography	α instead of αα	λ instead of λλ (medial): αλα
Dittography	αα instead of α	σσ instead of σ (medial): ερρωσσο
Crasis	crasis of α α	crasis of αι ε (word boundary): καιγω

Table 2. Overview of the description of phonological processes

With this terminology we can describe graphic variations, sometimes resulting from a mere slip of the pen, but more often from phonological changes in the Greek language, such as vowel length reduction, itacism, and the dropping of final -ς and -ν.[40] Diphthongs, double vowels, and double consonants are treated as a single unit to avoid splitting up the graphic realization of a single phoneme. The confusion of ει and ι is thus not represented as 'the addition of ε' but as 'ει instead of ι', and a haplography such as 'λ instead of λλ' is not described as 'the omission of λ'. This keeps these particular processes separate from other types of additions of vowels or from the omission of liquids in other environments.

Multiple irregularities in a single word are normally split up and described separately, even if they follow each other directly: rather than e.g. 'οδ instead of ωτ' we distinguish 'ο instead of ω' and 'δ instead of τ'. Strictly applying this principle could, however, represent morphosyntactic interchanges as being phonological in nature. An example is γίνεται which is

[40] For these linguistic processes see Gignac, *Grammar* 325, 235, 124–125, 111–112 respectively.

corrected by the editor to γίνονται: this could be represented as 'ε instead of ο' and 'omission of ν', but this clearly makes little sense although it is an accurate description of the change. It might be a better idea to refer to these morphological interchanges at the level of the morphemes, e.g. 'ε instead of ον' or 'εται instead of ονται'. In the same way, final -ω instead of -ον can be described as 'ω instead of ον' if the irregularity is interpreted as a morphosyntactic interchange of dative and accusative, but as 'ω instead of ο' and 'omission of ν (final)' if it is believed to be phonological in nature. In these situations a certain degree of interpretation is inevitable.

5.2. Providing possibilities for future research

Describing the editorial interventions at a phonological level does not automatically exclude interpretations of morphological and morphosyntactic variation. For instance, the correction of final -ου to -ω, although possibly phonological in nature, is likely to include many instances of the interchange of the genitive and the dative case. The correction of γυναικαν to γυναῖκα might not only be understood at the level of the addition of final -ν, but also in the light of the merger of inflectional paradigms. To facilitate access to these possible morphological and morphosyntactic interchanges, a parallel field for grammatical comments will be provided, containing e.g. 'nominative instead of genitive' or 'singular instead of plural'. This will allow for an analysis from different perspectives, enabling the study of processes with multiple causes and easy access to the grammatical phenomena.

Variation can be encountered at different levels of the language and editorial corrections contain many of these different types of variation.[41] This variety of examples in the database could be used to develop a typology for the different types of scribal variation. For this purpose, the additional field

[41] Schneider, in *Handbook of Language Variation* 67–96; H. Halla-aho, "Linguistic Varieties and Language Level in Latin Non-Literary Letters," in *The Language of the Papyri* 171–183.

for comments could also be used to comment on the type of variation, be it grammatical, graphical, lexical, or content-related. Furthermore, different categories could be distinguished for separate entities, such as personal names or numbers.

Editorial corrections are not only added to a text in order to 'correct' the scribe and to point out irregularities in the language. Editorial interventions were also meant as an aid for classically-trained scholars to read and understand the Greek used in documentary papyri. Whereas references to Classical Greek might be inappropriate when correcting the language of the scribe (see §4), they are suited to help understanding the meaning and grammar of a difficult word or phrase. When in *P.Oxy.* XIV 1683 (late IV) μαρτυρων (14) and λεβιτων (22) are corrected to μάρτυρα and λέβητα in the apparatus, this indicates 'understand accusative singular here'. In fact, the forms μαρτυρων and λεβιτων may have been meant as accusatives as well, with the interchange of ω and ο (very common in this text) and with the accusative singular morpheme (-ον) taken from the second instead of the third declension. Comments of this type, in this case preventing what at first sight looks like a genitive plural to be interpreted as such, will always be necessary in order to interpret the language correctly.

Editorial regularizations thus provide orthographic normalization and additional morphological categorization. As such they are an essential step towards a full lexicalization of the corpus of Greek papyri, complementing the Morpheus parsing and lemmatizing tool produced by the Perseus Project.[42]

5.3. Organizing flexible interaction between text and database

Establishing rules for the description of irregularities may be relatively easy because these annotations take place in a separate database environment. But the interaction between this stand-off database and the PN—or possible future databases from which corrections are mined—needs to be carefully modeled. No doubt the text edition environments need to

[42] http://wiki.digitalclassicist.org/Morpheus.

remain the place where corrections, regularizations, additions, and deletions of the ancient text are implemented and live in their most basic form. The database on the other hand could be the site where the type of correction is annotated and the type of variation specified, in the way described above. Detailed analyses such as 'final -αι following θ rendered as -ε' or free-form comments discussing alternative interpretations such as 'really confusion of cases or rather a phonetic variant?' should be possible.

To facilitate the interaction between full text corpus and stand-off database, two developments may be crucial. The first is a set of standards for the marking up of text with this type of editorial interventions. It will need to be decided, for example, which EPIDOC tags should be used in which circumstances. The second and probably more important issue is the establishment of a unique stable identifier that will allow the database to be informed about changes in the text editor, and the text editor to keep abreast of linguistic information. How this will work will need to be established, with attention for maximum flexibility and long-term stability. We invite whoever is interested in collaborating, either technically or linguistically, to contact us. Hopefully we will be able to set up a cooperative community around the linguistic study of the papyri, which will allow users to go much further in their analysis than we have done in this first tentative and exploratory paper.

November, 2014

Ancient History, KU Leuven
Blijde Inkomststraat 21 bus 3307
B-3000 Leuven, Belgium
mark.depauw@arts.kuleuven.be

Department of Philosophy, Classics,
History of Arts and Ideas
University of Oslo
Postboks 1020 Blindern
0315 Oslo, Norway
j.v.stolk@ifikk.uio.no